CONF-760428--1

LA-UR -76-661

TITLE: METHODS OF SIGNIFICANCE ARITHMETIC

AUTHOR(S): N. Metropolis

SUBMITTED TO: Conference Proceedings, "The State of the Art in Numerical Analysis"
University of York, England,
April 12-15, 1976.

los alamos
scientific laboratory
of the University of California
LOS ALAMOS, NEW MEXICO 87544

Form No. 836
St. No. 2629
1/75

UNITED STATES
ENERGY RESEARCH AND
DEVELOPMENT ADMINISTRATION
CONTRACT W-7405-ENG. 36

METHODS OF SIGNIFICANCE ARITHMETIC

N. Metropolis
Theoretical Division
Los Alamos Scientific Laboratory
Los Alamos, New Mexico 87545

2.1 Introduction.

The ambition of significance arithmetic is to be able to specify the standard deviations of computational results in the quite general case where input quantities have quite disparate magnitudes and accuracies. Many problems arising in the natural sciences are of this character, although too often their initial data are idealized by artificially extending their precision to that of the standard word.

A brief review of earlier developments: in its earliest form, significance arithmetic (SA) utilized the one degree of freedom in the computer representation of input quantities to exhibit their accuracy to the nearest integer value. A set of algorithms were found for addition (subtraction), multiplication and division that approximated the properties of appropriately combining statistically independent quantities. Exactly representable quantities were not distinguished; they were simply in normalized form. Later arithmetic rules were extended so that such precise quantities constituted a special set and each result was examined for truncation error and accordingly removed from the special set and henceforth regarded as an imprecise quantity. More recently, an axiomatic approach has been started based on equivalence classes of integer strings with a boundedness condition [1]. We do not pursue that direction here. Instead, the notion of non-integral values for the number of significant digits is introduced and developed.

It will be seen that this leads to a natural representation for the associated variance of a computer quantity, wherein its exponent in a suitably unnormalized form can be conveniently used as the exponent of the variance. This fact simplifies the physical realization in computing the variance of results of the fundamental arithmetical operations.

But for error correlations as a computation proceeds, the task of the error-analyst would be complete. The nature of these correlations is examined and analytic techniques developed for coping with them. For complicated and long sequences of arithmetical operations, a procedure is made available that reveals the correlations in an empirical fashion. A preliminary report is given of two applications--a simple nonlinear, partial differential equation and matrix inversion.

2.2 Non-integral values of significance and the role of variance.

The usual form of computer representation of un-normalized quantities is $x = ?^{e} \cdot f \equiv (e,f)$, for integer exponent e and the coefficient or fractional part f satisfying $0 \leq |f| < 1$. If x is an imprecise quantity, and σ_x is the associated standard deviation, define the number of significant digits of x as the rounded integer of

$$(2.2.1) \qquad s_x = \log \frac{|x|}{\sigma_x} .$$

Representation of x is not unique for $x = (e,f)$ is equal to $(e+a,\ 2^{-a}f)$ for integer a in, say, a binary computer. The range of a is such that no significant digits are lost on the right end of the

standard word† (and, of course, on the left). This one-degree of freedom may be used to exhibit the integer number of significant digits of x with reference to some fiducial bit position of f, call it k; it is usually near the right end of a standard word. The choice is optional, but once selected for a problem it is considered fixed. A quantity in this representation is said to be in significance form.

Assuming operands are statistically independent, one can establish rules for the representation of the result of addition (subtraction), multiplication and division of two quantities in significance, or unnormalized, form [2]. Such rules are, of course, limited to integer values of significant digits.

Somewhat later, provision was made to accommodate precisely representable operands [3], that were hitherto a source of possible awkwardness. Clearly if such an operand participates in an arithmetical operation with an imprecise one, the result is necessarily imprecise. If the participants of an operation are both precise, an examination is still required to establish whether the result is precise; the division 1 ÷ 3 illustrates the case of imprecise result from precise inputs.

The question naturally arose whether one could achieve a more accurate measure of significance by keeping track of the fractional part of s_x defined in (1). The idea would be to represent a quantity in significance form, i.e., with exponent and unnormalized fraction, together with a third part, the fractional significance. Considerations of how to manipulate these fractional significances in arithmetic processes made it clear that it would be more convenient to keep track of variances (in the sense of statistics). Specifically, the associated variance

---

† k is the residence of the least significant digit of x.

would be coupled to every operand; every arithmetic operation would then involve a corresponding computation of the variance of the result, with suitable approximations and simplifications described below.

2.3 Rules for variances.

Let $\langle x \rangle = E(x^2) - E^2(x)$ represent the variance of x, where E is expected value; $\{x,y\} = 2(E(x,y) - E(x)E(y))$ be the covariance of x,y. If x,y are statistically independent, then the rules from simple statistical considerations are:

(2.3.1) addition: $\langle x + y \rangle = \langle x \rangle + \langle y \rangle$

(2.3.2) multiplication: $\langle xy \rangle = x^2 \langle y \rangle + y^2 \langle x \rangle + \langle x \rangle \langle y \rangle$

(2.3.3) reciprocal: $\left\langle \frac{1}{x} \right\rangle \cong \frac{1}{x^4} \langle x \rangle$, (for $\langle x \rangle < x^2$)

where quantities outside the brackets are always mean values. The assumption is made that $\langle x \rangle < x^2$, otherwise x has no significant digits. Since $\langle x \rangle / x^2 = 2^{-2s_x}$, it is reasonable to neglect $\langle x \rangle \langle y \rangle$ in (2.3.2) no approximations are made in (2.3.1); the neglection in (2.3.3) is of order $\langle x \rangle^2 / x^6$. For division, the last two rules are combined in an obvious manner.

We remark that all numerical procedures should be monitored so that the computation is interrupted whenever a result has no significant digits. Note that in that circumstance, the associated variance is available for statistical considerations.

An important observation is that the above rules do not depend on the detailed structure of the distribution function associated with each

operand, apart from the natural assumption that the first and second moments exist.

2.4 Two representations.

Alternatives exist for the implementation of an arithmetic processor that would execute the coupled operations of the arithmetic proper and of the variance computation. In the first instance operands would be represented in unnormalized or significance form with appropriate algorithms [4]. In the second approach, the conventional normalized form of arithmetic is available.

The advantage of the former is that the exponent of $x = 2^e \cdot f$ is simply related to the exponent of its variance. (Recall that the representation of x is related to the magnitude of its variance.) As a consequence, the computer word need have only one exponent instead of two. Moreover, variance computation is more efficient in this form owing to fewer shift operations in the process of exponent matching. If one were embarking on a new design of an arithmetic processor, this approach should be seriously considered.

Normalized forms for arithmetic are, however, the more common approach and lend themselves to implementation by software. Here the couple operands would have their individual (not necessarily related) exponents along with their respective fractional parts and the pair of operations would be executed independently and presumably in new computers, concurrently.

Algorithms for computing the fractional parts of variances have been developed for both unnormalized and normalized forms of arithmetic;

the former has been implemented on the laboratory's MANIAC computer and the latter will be made available on one of the laboratory's commercial computers.

It is perhaps useful to consider a simple example, namely, that of summing a set of uncorrelated, imprecise operands, where no restrictions are placed on the magnitudes or imprecision of the individual summands. (The present writer is not aware of the existence of such a sub-routine in any other computer library.) If $S = \Sigma(x_i, \langle x_i \rangle)$ where $x_i = 2^e \cdot f_i$ is in significance form, then one orders the summands according to increasing exponent and adds them in turn accordingly, together with a calculation of successive variances. The ordering is desirable to avoid truncation on the right of the fractional part in a finite computer register.

On the other hand, if x_i is in normalized form, then it is the variances that are ordered and the summing of x_i is performed in that order.

2.5 Variance and significant digits

The variance of a result in addition, multiplication or division as a function of input quantities has a simple interpretation in terms of significant digits. Recall that $x^2/\langle x \rangle$ is related to the number of significant digits of x; in fact $2^{2s_x} = x^2/\langle x \rangle$ in binary base.

In the addition process, $\langle x + y \rangle = \langle x \rangle + \langle y \rangle$ for statistically independent x,y. Let

$$(2.5.1) \qquad c \frac{\langle x \rangle}{x^2} = \frac{\langle y \rangle}{y^2}, \qquad c > 0$$

that is, for $c < 1$, $s_y > s_x$; thus

$$(2.5.2) \qquad \langle x + y \rangle = \langle x \rangle \left(1 + c\frac{y^2}{x^2}\right).$$

The two terms in the parentheses determine the contributions of $\langle x \rangle$ and $\langle y \rangle$ respectively to the variance of the sum. If $c \approx 1$, and $x^2 \approx y^2$, $\langle x \rangle \approx \langle y \rangle$ and the variance of the sum has doubled, relative to a variance of the input. On the other hand, if $c < 1$ ($s_y > s_x$) and $y^2 < x^2$, then $\langle x \rangle$ contributes relatively more than $\langle y \rangle$ to $\langle x + y \rangle$. In significance arithmetic, this corresponds to $e_x > e_y$ where e_x, e_y are the exponents of x,y in significance form, and it is y that is shifted to the right to achieve exponnet match.

In multiplication, write Eq. (2b) as

$$(2.5.3) \qquad \frac{\langle xy \rangle}{x^2 y^2} = \frac{\langle x \rangle}{x^2} + \frac{\langle y \rangle}{y^2} + \frac{\langle x \rangle}{x^2} \cdot \frac{\langle y \rangle}{y^2}.$$

Since it is assumed that x,y have significant digits, the last term on the right in (2.5.3) can be neglected. Thus the number of significant digits of the product is determined primarily by the less significant input in accordance with the rules originally proposed for significance arithmetic [2]. For $s_x \approx s_y$, clearly $s_{xy} \approx s_x - \frac{1}{2}$.

Similarly, for division, a simple calculation shows that

$$(2.5.4) \qquad \frac{\langle x/y \rangle}{x^2/y^2} = \frac{\langle x \rangle}{x^2} + \frac{\langle y \rangle}{y^2} + \frac{\langle x \rangle}{x^2} \frac{\langle y \rangle}{y^2},$$

The comparison with Eq. (2.5.3) is interesting.

2.6 Error correlation.

If there is no error correlation between operands the above rules for computing variances would be reliable. Unfortunately, as a calculation proceeds, the operands tend to have correlated errors. In order to achieve reliable measures of variance, such correlations must be recognized and taken into account. In a simple algorithm, its tree structure may be examined in detail and a reliable version obtained. As experience increases and more sophisticated techniques are developed, more ambitious algorithms would become tractable. For very complicated algorithms, there exists an empirical statistical technique that can be applied to establish the extent of error correlation in the output. Moreover, if unacceptable amounts of correlation exist, then examination of intermediate quantities by the usual bisection of the program code pinpoints the source of correlation and can be dealt with. The technique is called the method of reduced precision and is discussed in the next section.

A simple example of error correlation occurs in forming $d = ab + ac$ where a,b,c are imprecise and statistically independent. Using either of the two arithmetics, one would find the appropriate value of $\langle d \rangle$ if the computation were performed as $d = a(b + c)$. On the other hand, if the computation was performed as $d = ab + ac$, then from statistics we know that

$$\langle d \rangle = \langle ab \rangle + \langle ac \rangle + \{ab, ac\}. \tag{2.6.1}$$

Since

(2.6.2) $\{ab,ac\} = bc\{a,a\} = 2bc\ \langle a \rangle$,

where, by our convention, quantities outside brackets are expected values. Thus the sum of the variances may not be a reliable measure of $\langle d \rangle$ because of error correlation. Note that the deviation may be of either sign according as the signs of b,c agree or disagree.

A second example of correlation (it occurs in the study of matrix inversion) is $\left\langle \frac{xy}{x+1} \right\rangle$ for x,y statistically independent. It can be shown that

(2.6.3) $$\left\langle \frac{xy}{x+1} \right\rangle = \left\langle \frac{y}{x+1} \right\rangle + \left(\frac{x-1}{x+1} \right) \langle y \rangle$$

and further simplification of $\langle y/(x+1) \rangle$ can be achieved using the easily derived relation $\langle x/y \rangle = \langle xy \rangle / y^2$. Some additional relations that may be useful in studying error correlations are given in the appendix.

2.7 Method of reduced precision

In complicated algorithms, the nature and detection of error correlation is less apparent. A method based on statistical perturbations of the initial data is available and one studies the consequent distributions of the output values. The natural setting for this method is the unnormalized form of operand representation that reveals the number of significant digits. Recall that each input has its least significant digit residing in the k^{th} stage on the right of a computer work, where k is optional but fixed throughout an algorithm. The idea is the following. Neglecting all correlations, one computes

a set of output (x_i^0) for the inputs (y_j^0) using significance arithmetic throughout, the superscript indicating original inputs and corresponding outputs; i and j take on the necessary range of values. The inputs are then statistically perturbed by adding a uniformly distributed random variable to the inputs so that in effect the precision of each is reduced by a constant amount, say three or four binary digits. The perturbed inputs have their least significant digit residing in the k' stage to the left of k. The inputs (y_j^1) produce (x_i^1). Form

$$(2.7.1) \quad \Delta_i^1 = x_i^1 - x_i^0 \qquad \text{for each i.}$$

Repeat the computation m times to achieve a distribution of values for Δ_i^m, for each i, starting with (y_j^m). If perturbed x_i is free of error correlation, then the distribution of Δ_i is strongly peaked about the k' + 1 position. More frequently, correlation does exist; for positive correlation the peak of the distribution is to the left of k' + 1, and to the right for negative correlation. In a given problem, all three possibilities may occur.

If error correlations are acceptably small, no further study is needed; otherwise two options are available. One can easily examine Δ-distributions for intermediate quantitites and pinpoint the source of correlation and then either modify the program code to eliminate that correlation, or to recognize its nature and use appropriate calculations of variances. The point is that the arithmetic processor would treat, for example, the (correlated) step q = x/(y+x) as though it were q' = x/(y+z). In this simple case one can, of course, write

$q = 1/(1 + (y/x))$ and avoid correlation. Clearly, there can be difficulties if one tacitly assumed $\langle q' \rangle = \langle q \rangle$. The perturbations of the initial data is to find such pitfalls.

This simple example of the two forms $x/(y+x)$ and $1/(1+(y/x))$ stimulates the remark that mathematically equivalent forms are not always computationally equivalent when dealing with imprecise quantities.

The method of reduced precision can, mutatis mutandis, also be used in normalized arithmetic. Significance arithmetic has a distinct advantage, however. Since its rules are based on uncorrelated errors, they approximate very closely the corresponding calculations of variances. Thus the method of reduced precision can dispense with such calculations until correlations are detected and needed. Normalized arithmetic must always include variance computation when using the method.

2.8 Two preliminary studies.

The coupling of variance computation with every arithmetic operation has been attempted in two instances: a study of Burgers' equation with initial and boundary conditions having imprecise values; a study of inversion of square matrices whose elements also have imprecise values with no restrictions on the disparity in magnitudes and imprecisions of such values. The studies are ambitious ones and we give only a preliminary report at this time.

Specifically, Burgers' equation in one space dimension is

$$(2.8.1) \quad u_t + uu_x - \nu u_{xx} = 0$$

where u has dimensions of velocity and the usual notation for (partial) space and time derivatives is adopted; and ν is the diffusivity coefficient. It is the simplest equation combining both nonlinear wave propagation and diffusive effects. The initial condition is a wave front dropping abruptly and continuously from uniformly high u-values behind the front to low u-values in front, also uniform. Boundary conditions at each end correspond to these uniform values respectively. Several discretized versions have been investigated; the simplest received most attention owing to the fact that the various interactions are complicated enough and the others did not offer any compensating advantages. Specifically, the difference equation is

$$(2.8.2) \quad u_j^{n+1} = u_j^n + \frac{\nu\Delta t}{\Delta x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) - \frac{\Delta t}{2\Delta x}\bar{u}(u_{j+1}^n - u_{j-1}^n)$$

where $\bar{u} = (u_{j+1}^n + u_j^n + u_{j-1}^n)/3$ and Δt, Δx are time and space intervals respectively and the superscripts and subscripts are time and space indices. The classical stability conditions relating Δt and Δx are of course observed.

Consider first the uniform region behind the wave front. Initially, statistical fluctuations about an equilibrium value are present; they correspond to a certain imprecision in the initial data. The method of reduced precision showed a decrease in the variances associated with u-values. What is the rate of decrease? Since the first and second space differences are essentially zero, the effect of continued processing of Eq. (12) at point j is to take sums of the initial, statistical fluctuations from a larger number of neighboring points surrounding the

the point j. This number increases linearly with n. The summing process is similar to the estimates, g_k, of the mean of a distribution; there it is well known that $g_k = \sigma^2/k$ where k is the size of the sample and σ^2 is the variance of the original distribution. Hence a measure of the decreasing variance of u_j^n in uniform regions is available.

A more interesting observation was the effect of the wave front passing through a region, wherein the u-values rise from some low value to some higher (uniform) one. In the region, say around point j, of the wave front, the variance of that u-value increases with time to a maximum and then decreases to its original value. Although the wave front being considered is not a shock wave, it is suggested that the increased variance be associated with an increase of entropy, a behavior well known in shock waves. If the variance of v is greater than that of u, the profile of the variances of u across the wave front exhibits a bimodal shape. The decrease at the center is connected with the fact that the second space difference of u has a minimum there. If the variance of v is small relative to that of u, the profile becomes unimodal and persists even as $\langle v \rangle \to 0$. Study is continuing to establish the nature of the dependence of $\langle u \rangle$ across the wave front. A likely candidate is the second space difference of u, possibly multiplied by u itself.

The second application is to inversion of a real, square matrix with elements of arbitrary magnitudes and significances. The discussion is limited to the 3 x 3 case. Our purpose is to study the nature of the correlations between the cofactors of the determinant in the algorithm of

interest. A tractable solution is found; it is clear how the analysis could be extended to 4 x 4 matrices, but the combinatorial complexity is great.

The algorithm studied is the classic one where the inverse element $a_{ij}^{-1} = a_{ji}^{*}/D$, where a_{ji}^{*} is the cofactor at j,i and D is the determinant expanded by either the i^{th} row or the j^{th} column. For each a_{ij}^{-1} there are six covariances to evaluate. If $a_{ji}^{*} = (uv + wx)$, two of the six are

$$(2.8.3)\quad \left\{\frac{u}{uv + wx}\,,\ \frac{v}{uv + wx}\right\},\ \left\{\frac{w}{uv + wx}\,,\ \frac{x}{uv + wx}\right\}$$

and four are permutations of

$$(2.8.4)\quad \left\{\frac{u}{uv + wx}\,,\ \frac{w}{uv + wx}\right\},$$

where the numerators select one member from each of the product pairs; uncorrelated factors have been suppressed.

The two covariances in (2.8.3) are exactly zero. Explicit expressions have been found for the remaining four given in (2.8.4) in terms of the individual variances of u,v,w,x. It turns out that if u,v,w,x are approximately of the same magnitude, then the covariances are small compared to variances and approach zero as equality is achieved. Finally $\langle a_{ij}^{-1} \rangle$ is given by the sum of these four covariances and the four variances, of which $\langle u/(uv + wx) \rangle$ is typical. (Further study may disclose that the covariances here make only small contribution to $\langle a_{ij}^{-1} \rangle$.)

Appendix

Some easily derived relations for statistically independent variables are:

1. $$\left\langle \frac{1}{x+1} \right\rangle = \left\langle \frac{x}{x+1} \right\rangle = \langle x \rangle / (x+1)^4$$

2. $$\left\langle \frac{1}{xy+z} \right\rangle = \frac{1}{(xy+z)^4} \left(\langle xy \rangle + \langle z \rangle \right)$$

3. $$\left\langle \frac{x}{xy+z} \right\rangle = \frac{1}{(xy+z)^4} \left(x^4 \langle y \rangle + \langle xz \rangle \right)$$

References

1. F. Faltin, N. Metropolis, B. Ross, and G.-C. Rota, The real numbers as a wreath product, Advances in Mathematics 16 (1975), 278-304.
2. N. Metropolis and R.L. Ashenhurst, Significant digit computer arithmetic, IRE Trans. Electron. Comput. EC-7 (1958), 265-267.
3. N. Metropolis, Analyzed binary computing, IEEE Trans. on Computers C-22 (1973), 573-576.
4. R.L. Ashenhurst and N. Metropolis, Unnormalized floating point arithmetic, Journ. ACM 6 (1959), 415-428.

This work was done in collaboration with R.L. Bivins and D. Wallwork; full accounts will be published elsewhere.